The GoldList Notebook Method: A Study on L2 Vocabulary Learning

John Duplice
Sophia University, Japan

Abstract

Vocabulary knowledge is of paramount importance when learning a second language. It requires effective and practical classroom vocabulary learning methods for long-term acquisition. Specific learning aspects helping the learner remember vocabulary such as spaced repetition and retrieval practice have shown efficacy but are often disconnected from practical in-class methods that can be used repeatedly. This study looked at data from 74 university students in Japan studying English vocabulary with the GoldList Notebook Method, which incorporates spaced learning and retrieval practice. The study was conducted over a nine-week period and consisted of a pre-test of target L2 idioms, a lesson teaching the idioms, the implementation of the GoldList Notebook Method, and a post-test on the target idioms. The data collected were analyzed through fixed effects with a generalized linear model in R version 4.0.3 and R Studio 1.2.5. In addition to quantitative data collected through the pre- and post-tests, qualitative observational data was compiled on the use of the GoldList Notebook Method in the classroom. The findings showed efficacy in using the method and found particular merit to spaced learning over two-week intervals. The study further addresses problems teachers may face implementing the method in the classroom and possible ways to alleviate the issues.

Keywords: desirable difficulties, L2 vocabulary, retrieval, spacing

The learning of a foreign language requires acquisition of many vocabulary words in the target language whether the learner is a beginner learning commonly used words or more advanced and learning specialized terms. With the advent of mobile applications such as Duolingo and Anki, vocabulary learning using pen-to-paper and notebooks has become less common. These applications can automatically provide spaced repetition based upon the user`s correct or incorrect responses identifying target vocabulary in a second language, henceforth referred to as L2. While they are commonly used for test-prep in memorizing of vocabulary, they often lack the context of target terms or the tactile connection in activities from pen-to-paper use. Furthermore, for learners who go beyond the dabbling stage of language learning, there is a need for a process requiring sufficient exposure and review of target L2 vocabulary (Nation, 2008; Schmitt & Schmitt, 2020). Additionally, the form and timing the L2 vocabulary exposure comes in is important. In Japan, students from middle school through university often study with the external primary purpose of passing tests such as school entrance exams. This often produces less than effective study habits for long-term retention and future use of L2 vocabulary in new learning, known as knowledge transfer. This focus on test-performance leads to use of methods that are less effective in long-term retention of L2 vocabulary.

The focus on test performance is described by Soderstrom and Bjork (2015) where *performance* is defined as "the temporary fluctuations in behavior or knowledge that can be observed and measured during or immediately after the acquisition process." (p.1) In comparison with performance, *learning* is the relatively permanent changes in knowledge where the knowledge acquired can be transferred to new learning (Soderstrom & Bjork, 2015). The aim of L2 vocabulary acquisition in the classroom should be the promotion of learning expressed as long-term retention over short-term performance. Desirable difficulties have been shown to help in learning over performance (e.g., Bjork, 1994; Bjork & Bjork, 2011; Brown, Roediger, & McDaniel, 2014; Persellin & Daniels, 2018). Desirable difficulties are activities or situations which increase the difficulty of a task initially. This difficulty slows the learning process down, requires increased focus, and often leads to less effective short-term performance results. Conversely, Ekuni, Vaz and Bueno (2011) explain that the increased difficulty can lead to better long-term recall and transfer ability, hence becoming desirable because the studying done was at a deeper processing level. It is important the difficulty be "desirable" and not so challenging the learner is unable to process the material or lose motivation to study. Two desirable difficulties commonly used in paper flashcards and mobile applications are spaced repetition and retrieval practice. Both of which are core aspects of an L2 vocabulary learning method known as The GoldList Notebook Method. Spaced repetition – repetition that has pre-set spaced time intervals where content, such as L2 vocabulary, is not reviewed until the next spaced interval – has been shown to increase long-term recall (Didau, 2015). Retrieval practice is an activity requiring the learner to retrieve material from memory. Spaced repetition and retrieval practice play upon one another as will be explained in more detail in the literature review section.

The fundamental concern of this paper is to explore a method of studying L2 vocabulary known as the GoldList Notebook Method among Japanese English language learners (ELLs) at university and its effect on the long-term retention of L2 vocabulary. The method incorporates spaced vocabulary recall practice through pen and paper rather than computer-based applications. Specifics on the rationale behind the GoldList Method are provided in the methodology section. The article reports on a study with the aim of identifying aspects of spacing out learning and retrieval practice in how they may enhance the long-term retention of L2 vocabulary among ELLs. While the focal point of the study is on university L2 learners in Japan, the GoldList Notebook Method itself may be incorporated into other classroom settings

with modifications. The modifications needed for all settings is beyond the scope of this article. Therefore, the article and study is limited to the EFL university classroom in Japan.

The GoldList Notebook Method incorporates retrieval and spacing desirable difficulties in the distillation process, where the learner attempts to recall the meaning of the vocabulary from memory. The study looks to answer whether the implementation of the GoldList Method as a supplement to a vocabulary lesson shows efficacy in increasing recall of vocabulary versus the vocabulary lesson alone. The study collected data from 74 first-year university students learning English at the Common European Framework (CEFR) B1 to B2 level. It looked at the merits of using the GoldList Notebook Method in learning L2 vocabulary in the form of idioms in a classroom and a variety of different spaced time-periods to glean insights to report to the teaching community. The paper first reviews the literature related to spacing and retrieval desirable difficulties in L2 learning and the pedagogical aspects of the GoldList Notebook Method. It then describes the methodology of the study, followed by a presentation of the findings and discussion before drawing conclusions and implications for teachers and learners of L2 vocabulary.

Literature Review

The role of desirable difficulties in L2 vocabulary acquisition has an established track record in effectively improving retention (Bjork & Kroll, 2015; Brown, Roediger & McDaniel, 2014; Didau, 2015). At the heart of the difference between performance and longer-term learning, is the difference between blocked practice and spacing out studying. Ebbinghaus (1885) as described in Didau (2015) illustrated there is a precipitous decline in what we remember soon after studying in the absence of time between review sessions. This lack of time is referred as blocked practice and is commonly used when cramming before a test and studying in a single block of time. In blocked practice, short-term performance is improved, hence the use of cramming. In spaced studying, short-term performance may be negatively affected, but longer-term learning is enhanced (Bjork, 1994; Glenberg, 1977; Kornell & Bjork, 2008; Randal, 2007; Rohrer, 2009). Bjork and Bjork (2011) explain the reason for the spacing effect as being the result of forgetting between study sessions and the need to subsequently review or relearn material leading to stronger connections to the target material. The active struggle of recalling material after forgetting it, is where the retrieval role comes in and is argued by Bjork (1975) that this process of forgetting and relearning through retrieval enables the learner to recall information more reliably in the future. The GoldList Notebook Method incorporates spacing to ensure against blocked practice by the learner. It also includes retrieval tasks following the spaced time away from the target vocabulary.

While there is much support for extrinsic activities such as retrieval in learning L2 vocabulary, there is another view that argues input through listening and reading alone are more effective. Krashen (1989) argues that study time should be focused on input in the form of reading and listening that is comprehensible. Krashen further states that language should be learned incidentally and subconsciously through natural acquisition. This stems from Krashen`s Monitor Model theory that argues there are two systems, an acquisition system and a learning system. According to Krashen, the acquisition system is the more important of the two for L2 acquisition. The learning system plays the role of editor or correcting mechanism and is only useful after language has been acquired (Krashen, 1989). Krashen argues that direct teaching of vocabulary and grammar should be avoided as the time is better spent receiving comprehensible input (Krashen, 1989; 1993). These arguments follow the idea that with enough exposure at the right level, vocabulary and grammar will be acquired naturally.

While Krashen argues to avoid explicit vocabulary instruction in the Second Language Acquisition (SLA) classroom, spacing and retrieval practice have been shown to provide support for learning target L2 vocabulary (Bird, 2011; Nakata, 2015; Nation, 2008; Sobel, Cepeda & Kepler, 2011). Schmitt (2010) argues the need for deeper connection or learning is of particular benefit in vocabulary retention as lexical knowledge is more apt to be forgotten through attrition than other linguistic aspects. The use of spaced repetition and retrieval are core aspects of mobile learning applications that focus on vocabulary acquisition. Teske (2017) argues an area that these applications may lack in efficacy, is that of teaching vocabulary in context. In L2 vocabulary learning, context is a core aspect of learning and may be the most important vocabulary learning strategy (Nation, 2008; Nation & Webb, 2011). With the GoldList Notebook, vocabulary can and should be introduced in context first through a lesson that could include reading, listening, writing, and speaking activities. The vocabulary would come from the lesson and be chosen by the instructor or the learner. Since the vocabulary comes from a specific lesson or activity, a connection between the target terms and the context they came from should be established before implementing the GoldList Method`s steps.

A major aspect of L2 vocabulary learning in the classroom and for research, is testing of knowledge. According to Nation and Webb (2011) and Schmitt (2010), vocabulary acquired in the short-term should not be considered learned. Post testing of at least three weeks should be used to evaluate whether the target material has been stored in long-term memory and to show with confidence whether target word acquisition is durable. For research purposes, it is also important to limit the exposure of the target vocabulary as to not influence the factors of retention (Nation & Webb, 2011; Schmitt, 2010). The GoldList Method allows for this amount of time, or more if needed, to ensure durability of vocabulary recall.

While the underlining activities such as spacing and retrieval practice have been shown to be effective in L2 vocabulary learning (Bjork & Kroll, 2015), the role of the GoldList Notebook Method in the L2 classroom has not been well researched. This method includes the same type of desirable difficulties as that of mobile applications, but also includes the tactile pen-to-paper aspect requiring the learning to write the target word increasing the amount of time and lexical attention. According to Schmitt (2010), this lexical attention and increased time interacting with the word help facilitate learning. Before implementing the GoldList Notebook Method as part of the L2 learning curriculum, empirical research needs to be done. Therefore, this action research study focuses on the GoldList Notebook Method`s efficacy in L2 vocabulary learning. Since this method can follow initial contextual learning and practice of the target vocabulary, spacing and retrieval aspects, and tactile pen-to-paper use, I hypothesize the method will show efficacy in the SLA university classroom. The present study attempts to answer the following questions:

1. Is the GoldList Notebook Method effective in moving L2 vocabulary from first exposure to long-term retention?
2. Is there a difference in efficacy among differing spaced time retrieval review periods of more than a week?
3. What takeaways are there for L2 language teachers to implement the GoldList Notebook Method into the classroom?

Methodology

Context and Rationale for the Study

An issue among students in Japan is the focus on performance over long-term learning. This is evident in the numerous entrance exams students take for each level of education. English is a core component of entrance exams and within the English part of the tests, vocabulary recognition is paramount for a high score. Students in Japan often focus their effort towards performance on these high-stake tests, over long-term learning and use of the language. This leads to blocked studying, known as cramming, where forced repetition through flashcard applications or rote memorization are often methods of choice. After passing the exam and entering university, students are left without an effective method of learning L2 vocabulary for the purpose of long-term recall and transfer. The GoldList Notebook Method is a method with some similarities to flash cards but with a few key differences. It requires the learner to first create a headlist of target vocabulary from the context of a lesson plan or other activity such as reading an article. It then has the learner space their exposure to the target vocabulary so there is time to forget before trying to retrieve the meaning or terms from memory. The process of forgetting and the subsequent struggle to recall have shown efficacy in longer-term retention (Soderstrom & Bjork, 2015). This takes place by increasing and strengthening neural pathway connections (Bjork, 1994; Bjork & Bjork, 2011), which is done in this study through distillations in the GoldList Notebook. What is not well established is the optimal time period to wait before a retrieval session. This study investigates different periods of time between distillations to see if there is a significant difference in long-term recall. The study also looks into the effectiveness of the GoldList Notebook Method as a whole post initial vocabulary lesson compared to the lesson alone. As this is an in-class action research study, observations beyond testing data are noted by the study`s facilitator that may be of importance to future studies or teachers looking to implement the GoldList Method.

Participants

In order to successfully answer the aforementioned questions, this study looked at data from 74 first-year ELLs attending a university in Tokyo, Japan. The participants were all Japanese L1 speakers, and the average age was 18. The participants were B1 to B2 CEFR English level as assessed by the Computerized Assessment System for English Communication (CASEC), which was conducted at the beginning of the school year by the university as an English level placement test. While the students were generally homogeneous in terms of background, their experience with English varied as some had lived abroad in high school. None of the students were English majors, so their primary English exposure came from the class this project studied. There were four classes with each constituting a group and forthwith will be referred to as Group. At the start of the study, the group sizes ranged from 17 to 24. Data from six participants was dropped from the study due to missing a quiz or the initial lesson. Therefore, data included in the results came from 74 participants with 21 males and 53 females. The four groups from which data was collected ranged from 16 to 22 participants per group (Group A = 16; Group B = 22; Group C = 19; Group D = 17). The groups were part of compulsory English courses that met for 100 minutes twice a week for fourteen weeks in the autumn semester from September 2021 to January 2022. The classes started online, being taught synchronously live via Zoom for the first five weeks before moving to face-to-face campus classes for the remaining nine weeks. English was the only language used by the instructor during the class sessions. Participants completed a consent form and were not provided compensation. Additionally, while the content in the study was part of the class, it was not considered in the

final grading and participants were informed at the beginning and throughout the study the pre- and post-tests would not be considered as part of their grade. The university where the study was conducted provided authorization, and participants were given the opportunity to have their data excluded at any time until the conclusion of the study. Risks such as possible data loss, were shared in the informed consent form and discussed prior to the start of the study.

The GoldList Notebook Method

The GoldList Notebook Method starts with a standard A4 sized paper notebook. This notebook replaces flashcards or mobile flashcard applications where target vocabulary is reviewed at pre-set intervals. The process of retrieving terms in the notebook is where the method plays a role in learning of vocabulary. It is not a lesson to introduce new vocabulary, but a tool for review and to strengthen recall. If the review sessions and spacing between the sessions are adhered to, the method provides the learner additional exposure to the target terms after the initial learning activity where the vocabulary was first introduced. This post-initial exposure is argued by Schmitt and Schmitt (2020) to be critical for consolidation. Furthermore, the method provides the desirable difficulties of spacing and retrieval practice which have been shown to increase retention of L2 vocabulary (Bjork & Kroll, 2015; Brown, Roediger, & McDaniel, 2014).

The process starts by laying the notebook flat with both the left and right page open. On each page a line is drawn down the middle vertically and a number is written on the left-hand column for each target vocabulary. A maximum of 20 words per session is recommended, but for more difficult terms or phrases fewer words may be more optimal. Ideally, the vocabulary used should come from an activity such as a reading task, lecture notes, or other lesson material where the terms are used in context and not chosen at random. The target word in L2 is written next to the number left of the line and the definition in L1, L2, or both is written on the right side of the line to the corresponding term. This initial list of vocabulary is known as the "headlist". Above the headlist, the date it was created and any contextual reference, including a title of where the terms originated from, is written. Once this is completed, the notebook is closed, and the terms are not looked at again until the first retrieval review, which is known as a distillation. At its core, a distillation is a self-quizzing activity to find out whether the learner recalls the meaning of the term. The amount of time between distillations is up to the learner, but 2 weeks is commonly used. During the first distillation, the learner covers the definitions and tries to recall the meaning of the target L2 term. If the term is unable to be recalled, a small checkmark is placed next to the word. The learner will then switch to the definition side of the headlist, cover the target vocabulary, and try to recall the word from the definitions. Here again, if it is not recalled, a check is placed next to the definition. At this point, any terms / definitions not recalled are written on the next page using the same format as the headlist but mixing up the order of the vocabulary. The order is mixed up to ensure the learner does not recall the word due to remembering the order. There are three of these distillations which follow a clockwise order with distillation 2 being written below distillation 1, and distillation 3 being written under the original headlist. Figure 1 below illustrates the process. Any words still not learned in distillation 3, can be put into a new headlist, moved to a separate notebook, or simply discarded if deemed unneeded by the learner.

Figure 1
GoldList Notebook Stages (Originated from Language Mentoring, n.d.)

Figure 2
GoldList Notebook Set from Headlist through Distillation 3. (Originated from Language Mentoring, n.d.)

Data Collection and Analysis

This study used quantitative data (Mayring, 2002) pertaining to recall of L2 vocabulary, but some observational qualitative data (Ivankova & Wingo, 2018) was used when assessing functionality and practical issues around the GoldList Notebooks use in the classroom setting.

The pre- and post-tests used to collect scores were conducted via Moodle learner management system quiz function. Condition and treatment were coded according to the initial fixed effects model. Baseline treatment was used as the baseline for comparison to scores from the other treatments. The lesson plan and tests were identical among all treatments with the only difference being the period between distillations. Dependent variables were determined by correct answers to both pre- and post-tests compiled as a mean score for each treatment. The GoldList Notebook activities were not conducted in the Baseline treatment in order to show efficacy of the other treatments versus the initial lesson without GoldList Notebook usage. The study used R version 4.0.3 and R Studio 1.2.5 to run the models. Additionally, packages included lme4 (Bates, Mächler, Bolker, & Walker, 2015) and pairwise (Heine, 2021).

Lesson Materials

The materials used for this study originated from ESL Library`s idiom lesson series titled "Detective Series 2: A Recipe for Disaster" (ESL Library, 2021). The first four lessons in the series were used and provided to the participants via pdf during class. Each lesson had between 9-11 idioms which were used as the target L2 vocabulary. The lessons also contained definition matching exercises, dialogue reading with target L2 idioms in context, synonym matching, and sentence generation activities. Idioms were used in place of single word vocabulary to better ensure participants did not have prior knowledge of the target L2 vocabulary. This way the baseline scores were more standardized and lower to better illustrate if the data were to show efficacy in the different treatments. During the setting up of the study, the tests were to be standard in the number of idioms tested of 10, but due to an error in coding the test in Moodle quiz function, the pre- and post-tests had a mean possible score of 9.75.

Procedure of the Study

The study consisted of four unique treatments: Frontloaded, Baseline, Mixed, and Standard as explained in the treatment descriptions and in Table 1. All groups were administered a pre-test for the target L2 idioms before being introduced in the lesson. The target idiom pre-test was in the form of a quiz with matching idioms to definitions in an answer bank. Time was limited to five minutes and participants were not allowed to use any type of reference material such as dictionaries or internet search engines during the tests. Following the pre-test, participants were provided the lesson and list of target idioms and given an hour to complete the lesson and review in groups of three to five participants. Each group had nine weeks between the initial pre-test, headlist, and lesson before taking the post-test, which was the same as the original pre-test. With the exception for the baseline treatment, each group had two weeks between the third distillation and the post-test, so there were no differences in length of time between final exposure to the target L2 idioms and the post-test except for the baseline treatment. Each group was exposed to the following treatments.

Treatments:

- Frontloaded treatment started with the headlist of idioms and focused on seeing if there was a benefit of retrieval session during the distillation task in weeks two, three, or seven during the use of the GoldList Notebook. Here the treatment focused on more exposure to the idioms early.
- Baseline treatment did not use the GoldList Notebook Method and simply had participants take the pre-test, complete the L2 target idiom lesson, and take the post-

test nine weeks later. The participants did not get any extra exposure to the target L2 idioms following the initial lesson.

- Mixed treatment had the participants do the first distillation in week two before waiting until week four for the second and week six for the third distillation.
- Standard treatment spaced all distillations exactly two weeks apart throughout the study taking place in weeks three, five, and seven. This differed from the other treatments that varied the time between distillations. Intervals of two weeks are the common length of time among polyglots utilizing the GoldList Notebook Method (Language Mentoring, n.d.).

Findings

Table 1 below is an example of the treatments and schedule. This example is of Group A. Other groups followed the same schedule, but the treatments alternated between groups so that each group was exposed to all the treatments once. This example is presented here to illustrate the distribution of treatments. Groups B, C, and D are similar, but the treatment to the sets varied.

Table 1
Example Schedule of Treatments from Group A

Task Date	Set 1 (Frontloaded)	Set 2 (Baseline)	Set 3 (Mixed)	Set 4 (Standard)
9/27	Headlist			
10/4	Distillation 1	Baseline		
10/11	Distillation 2		Headlist	
10/18			Distillation 1	Headlist
10/25				
11/1			Distillation 2	Distillation 1
11/8	Distillation 3			
11/15				Distillation 2
11/22			Distillation 3	
11/29	Post Test			Distillation 3
12/6		Post Test		
12/13			Post Test	
12/20				Post Test

Results of Pre- and Post-tests by Treatment

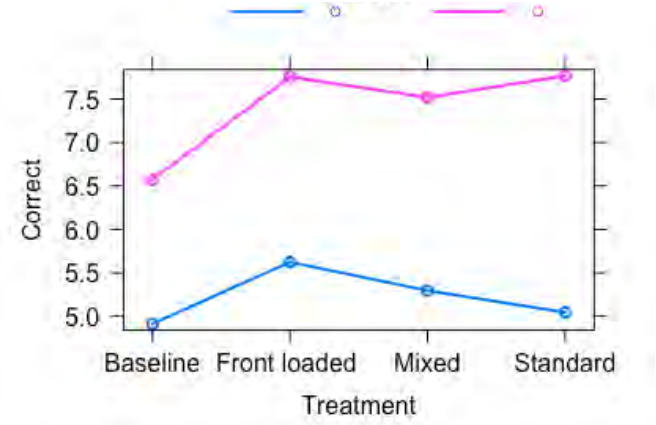
The results below in Table 2 show the pre- and post-test scores. Correct answers are shown as mean scores with standard deviation by treatment. Baseline treatment is used as the baseline compared with other treatment results for all participants.

Table 2
Descriptive Aggregate Results of Pre and Post Tests for Each Treatment in Mean (Standard Deviation)

Treatment	Pre-test	Post-test	t
Baseline	4.921	6.579	
	(1.958)	(2.424)	
Front loaded	5.632	7.763	2.316
	(2.006)	(2.052)	(1.092)
Mixed	5.303	7.526	1.244
	(2.239)	(2.138)	(1.304)
Standard	5.053	7.776	0.429
	(2.091)	(2.004)	(2.457)

Figure 3 provides a visual representation by test type and treatment. The mean for number of idioms tested among all four treatments is 9.75, but the figure shows only the actual mean scores from the tests.

Figure 3
Test Type and Treatment Effect Plot: Blue Line – Pre-Test Mean Scores: Pink Line - Post-Test Mean Scores.



An initial model included fixed effects of mean scores and treatment so that the pre-test was the baseline score, and the "Baseline" treatment was the baseline condition. The dependent variable was based on whether idioms were matched correctly with the definition in the pre- and post-tests. Therefore, a generalized linear model was used for fixed effects. As would be expected, the lack of review sessions in the Baseline treatment is reflected in it being the least effective. Comparing other treatments to the Baseline, the model showed significant interaction of the Standard treatment ($t = 2.457$, $p = .0143$).

Figure 3 above suggests an improvement to the model from all treatments when compared to the baseline, but only the Standard treatment showed significant improvement. The standard treatment included two weeks between each distillation, while the other treatments varied between one and three weeks throughout the study. The lmer package with difflsmeans function were used to conduct a pairwise comparison of treatments (exclusive of the fixed

comparison with the Baseline previously described), which confirmed the Standard treatment increased benefit over other treatments. In a nutshell, two weeks between review sessions throughout the study showed the greatest benefit with nearly twice the improvement of not using the GoldList Notebook Method. The other treatments also showed efficacy, but not to the same extent as the Standard two-week spacing between distillations. This provides evidence that spacing time between study sessions leads to better long-term recall of L2 vocabulary. In the following, more specifics will be provided directly relating to each of the three core questions asked at the beginning of the study.

Question 1: Is the GoldList Notebook Method effective in moving L2 vocabulary from first exposure to long-term retention?

This study looked at the use of the GoldList Method in learning L2 vocabulary specifically in the Japanese university classroom. The quantitative results showed increased learning of L2 idioms compared to not using the method. Therefore, the method shows efficacy over the lesson plan alone, referred to as the Baseline treatment in this study. This is described visually in Figure 3 where all the treatments show larger differences between pre-, and post-tests compared to the baseline. Specifically, the difference from baseline was most pronounced in the Standard treatment. The efficacy with that of the other treatments, was not as pronounced, but still showed improvement which is further discussed in Question 2.

Question 2: Is there a difference in efficacy among differing spaced time retrieval review periods of more than a week?

There was a difference in results between the three non-baseline treatments. The Standard treatment with two-week intervals between retrieval activities showed the most benefit. This was somewhat surprising as Ebbinghaus originally showed that spacing between study sessions when gradually expanded, is most effective (Didau, 2015). The benefit shown over the other treatments may be due to the consistency between study periods. Another possible reason is the extra week from initial idiom exposure in the lesson to the first distillation. Both the frontloaded and mixed treatments had only one week before the first distillation. This may indicate a need for more spacing of time after initial exposure to the target vocabulary. The two-week treatment may be optimal for class use as distillations can be scheduled regularly. Benefit of treatments is shown here from most beneficial to least: 1) Standard 2) Mixed 3) Front loaded 4) Baseline.

Question 3: What takeaways are there for L2 language teachers to implement the GoldList Notebook Method into the classroom?

Teachers of L2 who implement this method in their classroom should be aware students may forget their notebook at times. While not an option for most university classrooms, junior and high school settings where the same classroom is used daily, may allow for the notebooks to be kept in the classroom. This would alleviate the issue of students forgetting their notebook. This then allows for regular use from the headlist through the three distillations as long as students attend class on the days of the notebook is used. Notebooks could be given to students when complete if deemed necessary for future autonomous student use. It is also important for teachers to check notebooks regularly to ensure the activities are being done properly and regularly. In this study, the distillations were used as a class warm-up activity. This task allowed students the chance to retrieve L2 vocabulary and help clear external ruminations from smartphone use carried over from before the start of class. Educators have noted that

smartphone use is most distracting at the beginning of class, so an activity where smartphones are clearly not allowed helped in turning the students` attention towards the classroom content. Finally, early in the implementation of the study, many participants found the steps in the study unclear, so providing the students with a mentored example could be useful. In this study the facilitator worked on his own GoldList Notebook learning L2 Italian in front of the groups and provided example pictures in Italian on a projector screen of each step in the process.

Discussion

The present study revealed how effective the GoldList Notebook Method is when implemented in an L2 university classroom and whether some adjustments to the spaced time periods between distillations would show further benefit. The role of spacing and retrieval in the GoldList Notebook showed there is efficacy across all treatments and further benefit to longer spacing periods early in the review process as the initial two-week time was twice that of the other treatments. These findings confirm other findings in the literature on spacing (Bird, 2011; Nakata, 2015; Rohrer, 2009; Sobel, Cepeda, & Kapler, 2011) and retrieval practice (Bjork, 1975; Bjork & Kroll, 2015; Schmitt & Schmitt, 2020; Soderstrom & Bjork, 2015) as they relate to L2 vocabulary learning. During the lesson when the vocabulary items were introduced, participants received the most exposure through multiple tasks. These tasks included reading, matching, and generation activities through speaking, all of which were in context. The lesson was by far the most extensive exposure to the vocabulary the participants received, so their recall would likely be stronger soon after the lesson. This may have played a role in the increased initial struggle of the Standard two-week spaced gap versus the one-week of the other treatments. As described by Soderstrom & Bjork (2015) and Bjork (1994), this extra difficulty likely caused more difficulty in recalling terms and therefore making stronger connections, which developed into better recall later in the Week Nine post-test. Additionally, the consistency of how much time between distillations of the Standard treatment may have played a role in the increased efficacy. This made it easier for participants to schedule their distillations and less likely to mistake it with another distillation from a different lesson.

Finally, the qualitative observations made throughout the study provide insights into possible pitfalls to implementing the method in the classroom. Teachers should be aware that students in university courses, and therefore likely in pre-university classes as well, have issues with forgetting to bring the notebook to class. Additionally, they frequently omitted the date at the top of each distillation. One possible solution to this issue is keeping the notebooks in the classroom where the teacher has more control to ensure the notebooks are available. However, this may not be feasible for all settings such as the university setting this study was set in. By not allowing the students access to the notebooks between distillations, extra exposure to the terms is further limited thus leading to more struggle to recall during distillations and therefore more durable neural connections. An option when keeping the notebooks in class is not possible, is assigning students to take a picture and upload it to a learner management system on the day the distillation is due. This will not limit the possibility of extra exposure by eager students, but it can help in confirming students are completing the distillations as scheduled. Finally, use of size A4 notebook is recommended for a couple of reasons. First, smaller notebooks do not allow enough space for distillation 3 if the original headlist is 20 words. There needs to be enough space under the headlist in case there are more than a couple terms still not known at this stage. Likewise, skipping lines between headlist terms also limits space for the third distillation. The A4 size enables students to include more than just translated definitions. If there is enough room, example sentences, drawings, or L2 synonyms can also be included in the headlist further adding possibilities for contextual recall.

Limitations and Suggestions for Future Research

This study investigated a vocabulary learning method with little published research in general and particularly in the classroom. The study began virtually, making oversight of adherence by participants more difficult. The study was also limited in scope to the specific three questions in the discussion section. While the implementation of a L2 vocabulary learning application may be just as or more effective as the GoldList Notebook findings, this study did not directly compare the two. The number of idioms used in the study may have been too lengthy or not long enough for some participants. Due to the Covid 19 situation at the end of the semester, the study was not able to implement a participant questionnaire for feedback which may have provided more insight into the student perspectives and specifics such as number of items used in each set.

Another limitation arose from participants forgetting to bring the GoldList Notebook to class. As the class started online, this issue was not initially apparent. It is not clear if the participants were actually writing in the notebook or on a separate sheet of paper and later putting the information into the notebook, which could affect the data results. This action would provide the participants an extra exposure and eliminate the specified spaced time between distillations for each treatment. In an attempt to solve this issue, participants were required to present the GoldList Notebook in class for the instructor to check. If numerous distillations were missed or a notebook was lost, the participant data was omitted from the study, hence 6 participants were dropped from the results.

Another issue that arose was the lack of writing dates in the notebook for the headlists and distillations. Since distillations were conducted live in class and facilitated by the instructor, this did not affect the data. For more autonomous use of the GoldList Notebook Method, the lack of dates could become an issue. Additionally, while participants were instructed to not write the missed target L2 idioms when performing distillations in the same order as a previous distillation or headlist, many participants did not comply.

Conclusion

This study tested the GoldList Notebook Method and the impact of spacing and retrieval integrated into the activities on the long-term recall of L2 vocabulary among university students in Japan. The findings contribute to teachers looking for a low-stakes and inexpensive way to regularly expose students to target vocabulary without having to use daily quizzes which may cause anxiety among the learners. It is significant to state that the GoldList Notebook Method is promising for teens and adults but may be inappropriate for use with very young ELLs or ELLs who are newcomers arriving from harsh places.

The research conducted in this study shows that there is benefit through making small additions or adjustments to the curriculum. Much can be achieved in as little as a few minutes at the beginning of class as was the case here. Spacing and retrieval are strategies that can be implemented into many different parts of the L2 curriculum. Having a simple method teachers and students can use with confidence, is very practical. The teacher-researcher knows best what can be easily implemented into the curriculum and whether the GoldList Notebook Method might work in his/her class. In future research, investigation into the difference between mobile applications and the GoldList Notebook Method is warranted. If similar benefit can be achieved with mobile applications, facilitation of L2 vocabulary review among students may be more

easily conducted by the teacher. The GoldList Notebook Method is just one tool for teachers to consider, implement, experiment with, and adjust to their learners` needs.

References

Bates, D., Mächler, M., Bolker, B., & Walker, S. (2015). Fitting linear mixed-effects models using lme4. *Journal of Statistical Software*, *67*(1), 1–48. https://doi.org 10.18637/jss.v067.i01

Bird, S. (2011). Effects of distributed practice on the acquisition of second language English syntax—erratum. *Applied Psycholinguistics*, *32*(2), 435–452. https://doi.org/10.1017/S0142716410000172

Bjork, R. A. (1975). Retrieval as a memory modifier: An interpretation of negative recency and related phenomena.

Bjork, E. L., & Bjork, R. A. (2011). Making things hard on yourself, but in a good way: Creating desirable difficulties to enhance learning. *Psychology and the Real World: Essays Illustrating Fundamental Contributions to Society*, *2*(59–68).

Bjork, R. A. (1994). Memory and metamemory considerations in the training of human beings. *Metacognition: Knowing about Knowing*, *185*(7.2).

Bjork, R. A., & Kroll, J. F. (2015). Desirable difficulties in vocabulary learning. *The American Journal of Psychology*, *128*(2), 241–252. https://doi.org/10.5406/amerjpsyc.128.2.0241

Brown, P. C., Roediger III, H. L., & McDaniel, M. A. (2014). *Make it stick: The science of successful learnin*g. Harvard University Press. https://doi.org/10.4159/9780674419377

Didau, D. (2015). *What if everything you knew about education was wrong?* Crown House Publishing.

Ekuni, R., Vaz, L. J., & Bueno, O. F. A. (2011). Levels of processing: The evolution of a framework. *Psychology & Neuroscience*, *4*(3), 333–339. https://doi.org/10.3922/j.psns.2011.3.006

ESL Library. (2021) Detective series 2: A recipe for disaster. *ESLlibrary.com*

Glenberg, A. M. (1977). Influences of retrieval processes on the spacing effect in free recall. *Journal of Experimental Psychology: Human Learning and Memory*, *3*(3), 282. https://doi.org/10.1037/0278-7393.3.3.282

Heine, J. (2021). *Pairwise: Rasch model parameters by pairwise algorithm*. R package version 0.5.0-2. https://CRAN.R-project.org/package=pairwise.

Ivankova, N., & Wingo, N. (2018). Applying mixed methods in action research: Methodological potentials and advantages. *American Behavioral Scientist*, *62*(7), 978–997. https://doi.org/10.1177/0002764218772673.

Kornell, N., & Bjork, R. A. (2008). Learning concepts and categories: Is spacing the "enemy of induction"? *Psychological Science*, *19*(6), 585–592. https://doi.org/10.1111/j.1467-9280.2008.02127.x

Krashen, S. D. (1989). We acquire vocabulary and spelling by reading: Additional evidence for the input hypothesis. *The Modern Language Journal, 73*(4): 440–464. https://doi.org/10.1111/j.1540-4781.1989.tb05325.

Krashen, S. D. (1993). The effect of formal grammar teaching: Still peripheral. *Tesol Quarterly*, *27*(4), 722–725. https://doi.org/10.2307/3587405

Language Mentoring. (n.d.). *The Goldlist Method: Learning vocabulary has never been easier!* https://www.languagementoring.com/goldlist-method-how-to-learn-vocabulary/

Mayring, P. (2001, February). Combination and integration of qualitative and quantitative analysis. *In Forum Qualitative Sozialforschung/Forum: Qualitative Social Research, 2*(1), https://doi.org/10.17169/fqs-2.1.967

Nakata, T. (2015). Effects of expanding and equal spacing on second language vocabulary learning: Does gradually increasing spacing increase vocabulary learning? *Studies in Second Language Acquisition*, *37*(4), 677–711 https://doi.org/10.1017/S0272263114000825

Nation, I. S. (2008). *Teaching ESL/EFL reading and writing*. Routledge. https://doi.org/10.4324/9780203891643

Nation, I. S., & Webb, S. A. (2011). *Researching and analyzing vocabulary*. Heinle, Cengage Learning.

Persellin, D. C., & Daniels, M. B. (2018). *A concise guide to teaching with desirable difficulties*. Stylus Publishing, LLC.

Randall, M. (2007). *Memory, psychology and second language learning* (Vol. 19). John Benjamins Publishing. https://doi.org/10.1075/lllt.19

Rohrer, D. (2009). Research commentary: The effects of spacing and mixing practice problems. *Journal for Research in Mathematics Education*, *40*(1), 4–17. https://doi.org/10.1007/s11251-007-9015

Schmitt, N. (2010). *Researching vocabulary: A vocabulary research manual*. Springer.

Schmitt, N., & Schmitt, D. (2020). *Vocabulary in language teaching*. Cambridge University Press. https://doi.org/10.1017/9781108569057

Sobel, H. S., Cepeda, N. J., & Kapler, I. V. (2011). Spacing effects in real-world classroom vocabulary learning. *Applied Cognitive Psychology*, *25*(5), 763–767. https://doi.org/10.1002/acp.1747

Soderstrom N. C., Bjork, R. A. (2015) Learning versus performance: An integrative review. *Perspectives on Psychological Science*,10(2):176-199. https://doi.org/10.1177/1745691615569000

Teske, K. (2017). Duolingo. *Calico Journal*, *34*(3), 393–401. https://doi.org/10.1558/cj.32509

Corresponding author: John Duplice
Email: duplice@outlook.com